The Annals of Statistics
1989, Vol. 17, No. 3, 1242–1256

ROBUST NONPARAMETRIC REGRESSION ESTIMATION FOR DEPENDENT OBSERVATIONS

BY GRACIELA BOENTE AND RICARDO FRAIMAN

Conicet and Universidad de Buenos Aires, and Universidad de Buenos Aires

Robust nonparametric estimators for regression and autoregression are proposed for φ- and α-mixing processes. Two families of M-type robust equivariant estimators are considered: (i) estimators based on kernel methods and (ii) estimators based on k-nearest neighbor kernel methods. Strong consistency of both families is proved under mild conditions. For the first class the result is true under no assumptions whatsoever on the distribution of the observations.

1. Introduction. There are many nonparametric methods for estimating the regression function in the i.i.d. case [see for instance Collomb (1981) for a review], some of which have been extended to time series models.

Two of the most common methods in nonparametric regression are kernel and k-nearest neighbor kernel methods, introduced by Nadaraya (1964) and Watson (1964) and by Collomb (1981), respectively. Both of them are weighted averages of the response variables and therefore are highly sensitive to large fluctuations in the data. Thus, these estimates are not asymptotically qualitatively robust as defined for stochastic processes in Papantoni-Kazakos and Gray (1979), Cox (1981) or Boente, Fraiman and Yohai (1987) where the concept of asymptotically strongly robust (ASR) is introduced. Robust estimators can be obtained via M-estimates. A first approach in the i.i.d. case was given in Tsybakov (1982) and Härdle (1984), who studied pointwise asymptotic properties of a robust version of the Nadaraya–Watson method when scale is known. Later on, Härdle and Tsybakov (1988) extended their previous results to M-type scale equivariant kernel estimates. See also Boente and Fraiman (1989), who consider robust scale equivariant nonparametric estimates using nearest neighbor weights and weights based on kernel methods by applying a robust location functional to estimates of the conditional empirical distribution function of the response variables.

In this paper we study the case when the observed sample has serial dependence.

Let $\{(X_t, Y_t): t \geq p+1\}$ be a strictly stationary process, $X_t \in \mathbb{R}^p$ and $Y_t \in \mathbb{R}$. For $x \in \mathbb{R}^p$ let $\phi(x) = E(Y_t | X_t = x)$. The Nadaraya–Watson regression estimator defined by

$$\phi_T(x) = \sum_{t=p+1}^{T} w_{tT}(x) Y_t, \tag{1.1}$$

Received January 1986; revised October 1988.

AMS 1980 *subject classifications*. Primary 62G05; secondary 62M10.

Key words and phrases. Robust regression estimation, kernel estimation, k-nearest neighbor estimation, φ-mixing, α-mixing, strong consistency.

where

$$w_{tT}(x) = w_{tT}(x, X_{p+1}, \ldots, X_T) = K((X_t - x)/h_T) \Big/ \sum_{\tau=p+1}^{T} K((X_\tau - x)/h_T),$$

K is a nonnegative integrable function on $\mathbb{R}^p$ and $h_T > 0$, was applied by Watson (1964) to a meteorological prediction problem. Later Roussas (1969), Bosq (1980), Doukhan and Ghindes (1980, 1983), Collomb (1982, 1984), Robinson (1983), Yakowitz (1985) and Doukhan, Leon and Portal (1985) studied asymptotic properties of such estimators and predictors. In particular, these estimators are used for a pth order autoregressive model, i.e., a strictly stationary real valued process $\{Z_t: t \in \mathbb{N}\}$ satisfying

$$Z_t = g(X_t) + e_t, \tag{1.2}$$

where $X_t = (Z_{t-1}, \ldots, Z_{t-p})$, $Y_t = Z_t$, e_t is independent of $\{Z_{t-1}, Z_{t-2}, \ldots\}$ and $E(e_t) = 0$.

We consider processes $\{(X_t, Y_t): t \geq p + 1\}$ not necessarily defined from $\{Z_t: t \in N\}$ and in this way we also include the i.i.d. case, although we always have in mind a pth order autoregressive model. This framework also includes the transfer function models of Box and Jenkins (1970).

As noted above, $\phi_T(x)$ is a type of weighted average of the observations $\{Y_t: t \geq p + 1\}$ and can be viewed as minimizing the quadratic loss function $\sum_t(Y_t - \phi_T(x))^2 w_{tT}(x)$. Evidently, $\phi_T(x)$ is highly sensitive to the effect of just one isolated disparate observation Y_t, particularly if the corresponding X_t is close to x. Note also that strong consistency results for these linear methods require that $|Y_t| \leq M < \infty$ for all t; see Collomb (1984) or Peligrad (1986).

Robinson (1984) adapted robust M-estimators of a location parameter with kernel weights to time series models replacing the quadratic loss function by a loss function related to a convex function ρ with bounded derivative ψ, and established a central limit theorem for such estimators when scale is known. A similar approach was considered by Collomb and Härdle (1984) who established uniform convergence of this family of estimators for φ-mixing processes. In this paper, we consider robust scale equivariant nonparametric M-estimators based not only on kernel methods but also on k-nearest neighbor kernel methods for which we obtain strong pointwise convergence under quite mild conditions. We follow the approach developed in Boente and Fraiman (1989).

More precisely, let (X, Y) be a random vector with the same distribution as (X_t, Y_t). The robust conditional location functional $g(X) = E^{\psi}(Y|X)$ defined in Boente and Fraiman (1989) is the essentially unique $\sigma(X)$-measurable function $g(X)$ that verifies

$$E\{h(X)\psi[(Y - g(X))/s(X)]\} = 0 \tag{1.3}$$

for all integrable function h, where $\sigma(X)$ is the σ-algebra generated by X, $s(X)$ is a robust measure of the conditional scale, e.g.,

$$s(x) = \text{med}(|Y - m(x)|) = MAD_c(x), \tag{1.4}$$

$m(x) = \text{med}(Y|X = x)$ is the median of a regular version $F(y|X = x)$ of the conditional distribution function and $\psi: \mathbb{R} \to \mathbb{R}$ is a strictly increasing, bounded and continuous function. When the distribution of $Y|X = x$ has half or more than half of its mass at one single point we redefine $s(x) = 1$. If the conditional distribution function $F(y|X = x)$ is symmetric around $\phi(x)$ and ψ is odd, we have $g(x) = \phi(x)$. Then, in this sense it is a natural extension of the conditional expectation $E(Y|X)$.

In Theorem 2.1 of Boente and Fraiman (1989) it was shown that the solution of (1.3) exists, is unique and measurable. The weak continuity of the functional so defined was proved in Theorem 2.2. Then, we obtain consistent and asymptotically strongly robust estimates of the autoregression function by applying this functional to estimates $F_T(y|X = x)$ of $F(y|X = x)$, verifying that $F_T(y|X = x) \to_\omega F(y|X = x)$ as $T \to \infty$ a.s. (μ), where $\to_\omega$ stands for weak convergence and μ denotes the marginal distribution of the vector X.

We will consider two families of estimators of $F(y|X = x)$.

1. Estimators based on kernel weights. These are defined by

$$F_T(y|X = x) = \sum_{t=p+1}^{T} w_{tT}(x) I_A(Y_t), \tag{1.5}$$

where $A = (-\infty, y]$, I_A denotes the indicator function of the set A and w_{tT} is defined in (1.1).

2. Estimators based on k-nearest neighbor kernel methods. These are defined by

$$\hat{F}_T(y|X = x) = \sum_{t=p+1}^{T} \hat{w}_{tT}(x) I_A(Y_t), \tag{1.6}$$

where

$$\begin{aligned} \hat{w}_{tT}(x) &= \hat{w}_{tT}(x, X_{p+1}, \dots, X_T) \\ &= K((X_t - x)/H_T) \Big/ \sum_{\tau=p+1}^{T} K((X_\tau - x)/H_T), \end{aligned}$$

H_T is the distance between x and the k-nearest of x among $X_{p+1}, \dots, X_T$ and $k = k_T$ is a fixed integer. In particular, when $K(t) = I_{\{\|u\|<1\}}(t)$, where $\|\cdot\|$ is any norm on $\mathbb{R}^p$, we obtain the uniform k-NN estimate.

Denote by $s_T(x)$ and $\hat{s}_T(x)$ the scale measures corresponding to $F_T(y|X = x)$ and $\hat{F}_T(y|X = x)$, respectively, as defined in (1.4). The corresponding robust nonparameteric estimates of $g(x)$ are given by the unique solutions of

$$\sum_{t=p+1}^{T} w_{tT}(x) \psi[(Y_t - g_T(x))/s_T(x)] = 0 \tag{1.7}$$

and

$$\sum_{t=p+1}^{T} \hat{w}_{tT}(x) \psi[(Y_t - \hat{g}_T(x))/\hat{s}_T(x)] = 0, \tag{1.8}$$

respectively.

We obtain strong pointwise consistency of both families of estimates, when the observations (X_t, Y_t) have a φ- or an α-mixing dependence structure [see Billingsley (1968) and Rosenblatt (1956), respectively]. For kernel weights, the consistency results hold without requiring any regularity condition to the distribution of the process. In this sense they can be considered distribution-free results. For k-nearest neighbor kernel weights, it is required that the vector X have a density $f(x)$.

In Sections 2 and 3, assumptions and strong pointwise consistency results are stated for kernel and k-nearest neighbor kernel weights, respectively. In Section 4, two useful exponential inequalities for φ- and α-mixing processes, respectively, are stated and the almost everywhere weak convergence of the conditional distribution functions defined in (1.5) and (1.6) is proved.

As is well known the concept of α-mixing (strongly mixing) processes includes φ-mixing (uniform strongly mixing) processes. The φ-mixing condition is rather restrictive when we are considering autoregressive models while α-mixing processes generally include a pth order autoregressive model as defined in (1.2). However, we consider both cases separately. In the φ-mixing case we obtain a better convergence rate on the window bandwidth h_T than for strong mixing processes. We also do not require a geometric condition on the mixing coefficients for φ-mixing processes. Some examples of possible models under consideration are given in Section 2.

2. Asymptotic results for robust nonparametric estimates based on kernel methods. We will consider the following set of assumptions.

A1. $\psi\colon R \to R$ is a strictly increasing, bounded and continuous function such that $\lim_{u\to\infty} \psi(u) = a > 0$ and $\lim_{u\to-\infty} \psi(u) = b < 0$.

A2. Either of the following statements holds.

(a) $s(x)$ is given by a functional which is weakly continuous at F, for almost all x.

(b) ψ is odd and $F(y|X = x)$ is symmetric around $g(x)$ and a continuous function of y for each fixed x.

H1. Either of the following statements holds.

(a) The process $\{(X_t, Y_t)\colon t \geq p + 1\}$ is a strictly stationary φ-mixing process (uniform strongly mixing), i.e. [Billingsley (1968), page 166], there exists a nonincreasing sequence of positive numbers $\{\varphi(n)\colon n \in N\}$, with

$$\lim_{n\to\infty} \varphi(n) = 0$$

such that for any integer n,

$$|P(A \cap B) - P(A)P(B)| \leq \varphi(n)P(A),$$

where $A \in M_{p+1}^{t}$, $B \in M_{t+n}^{\infty}$ and M_u^v is the σ-field generated by the random vectors $\{(X_t, Y_t)\colon u \leq t \leq v\}$.

(b) The process $\{(X_t, Y_t): t \geq p+1\}$ is an α-mixing process (strongly mixing), i.e. [Rosenblatt (1956)], there exists a nonincreasing sequence of numbers $\{\alpha(n): n \in N\}$ with $\lim_{n\to\infty} \alpha(n) = 0$ such that for any integer n,

$$|P(A \cap B) - P(A)P(B)| \leq \alpha(n),$$

where $A \in M_{p+1}^{t}$ and $B \in M_{t+n}^{\infty}$. We also assume that the mixing coefficients are geometric, i.e., there exist $0 < \rho < 1$ and $a > 0$ such that $\alpha(n) \leq a\rho^n$.

Remark 2.1. The φ-mixing condition is rather restrictive when we consider autoregressive processes, for instance a stationary gaussian process is φ-mixing if and only if it is m dependent for some $m < \infty$ [see Ibragimov and Linnik (1971), Theorem 17.3.2], as was noted by Collomb and Härdle (1984), where they suggest that a natural context should be the case of strong mixing processes. As is well known, the class of strong mixing processes (α-mixing) generally includes the pth-degree autoregressive model defined by (1.2). However, in the φ-mixing case we obtain a better convergence rate on the window bandwidth h_T than for the strong mixing processes and we do not require a geometric mixing condition as in the α-mixing case.

H2. $K: R^p \to R$ is a bounded nonnegative function satisfying

$$aI_{\{\|u\| \leq r\}}(u) \leq K(u) \quad \text{for some } a > 0,\ r > 0,$$

$$a_1 H(\|u\|) \leq K(u) \leq a_2 H(\|u\|),$$

where a_1 and a_2 are positive numbers and $H: R^+ \to R^+$ is bounded, decreasing and such that $t^p H(t) \to 0$ as $t \to \infty$.

H3. The sequence $\{h_T: t \in N\}$ is such that

$$h_T \to 0 \quad \text{and} \quad Th_T^p \to \infty \quad \text{as } T \to \infty.$$

H4. (a) When we are dealing with φ-mixing processes we require the sequences $\{h_T: T \in N\}$ and $\{\varphi(n): n \in \mathbb{N}\}$ to satisfy the following: There exist $A > 0$ and a nondecreasing sequence $\{n_T: T \in N\}$ such that

$$1 \leq n_T \leq T, \qquad T\varphi(n_T)/n_T \leq A$$

and

$$Th_T^p/(n_T \log T) \to \infty \quad \text{as } T \to \infty.$$

(b) In the α-mixing situation we will use: There exists $\delta > 0$ such that $T^{1/4}(h_T^p)^{(1+\delta)/4}/\log T \to \infty$ as $T \to \infty$.

Remark 2.2. (a) If $\{Z_t: t \in N\}$ is a strictly stationary φ- or α-mixing process, then $\{(X_t, Y_t): t \geq p+1\}$ is also a φ- or α-mixing process, respectively.

(b) Assumption H4(a) has been introduced by Collomb (1984) and as was noted in his paper, H4(a) is fulfilled in the following three interesting situations:

(i) The process $\{(X_t, Y_t): t \geq p+1\}$ is m-dependent [$\varphi(n) = 0$ for $n > m$]. In this case what is required is just $Th_T^p/\log T \to \infty$ as $T \to \infty$.

(ii) $\{(X_t, Y_t): t \geq p + 1\}$ is a geometrically φ-mixing process, i.e., there exist $0 < \rho < 1$ and $a > 0$ such that $\varphi(n) \leq a\rho^n$. If we choose $n_T = c \log T$, where $c > -1/\log \rho$, then the required rate is $Th_T^p/\log^2 T \to \infty$ as $T \to \infty$.

(iii) There exist $w > 1$ and $a > 0$ such that $\varphi(n) \leq an^{-w}$. In this case we are requiring that $Th_T^p/(T^{1/(1+w)} \log T) \to \infty$ as $T \to \infty$.

REMARK 2.3. All the consistency results can also be obtained, without requiring that the α-mixing coefficients have geometric behaviour, by using an exponential inequality due to Carbon (1983) instead of Lemma 5.2. However, Lemma 5.2 gives a weaker condition on the bandwidth selection. More precisely, Carbon's inequality requires that there exists a sequence $\{n_T: T \in \mathbb{N}\}$, $1 \leq n_T \leq T$ for which the kernel bandwidth satisfies $Th_T^p/(n_T \log T) \to \infty$, $T\tilde{\alpha}(n_T)/n_T^2 \leq A$ and $T\alpha(n_T)^{2n_T/5T}n_T^{-1} \leq B$, where A and B are positive constants and $\tilde{\alpha}(n) = \sum_{i=1}^{n} \alpha(i)$. In particular, in a geometric α-mixing process h_T must satisfy the stronger requirement $T^{(1-\beta)/2}h_T^p/\log T \to \infty$ for some $0 < \beta < 1$.

We will need the following lemma which can be found in Boente and Fraiman (1989), Theorem 2.2.

LEMMA 2.1. *Assume* A1 *and* A2. *Let* $\{F_n(x, y)\}$ *be a sequence of distribution functions such that*

$$F_n(y|X = x) \to_\omega F(y|X = x) \quad a.s.\,(\mu),$$

where $F_n(y|X = x)$ *stands for the conditional distribution function of* $F_n(x, y)$.
Then we have that

$$E_{F_n}^{\psi}(Y|X) \to E_F^{\psi}(y|X) \quad a.s.\,(\mu),$$

where by $E_G^{\psi}(y|X)$ *we denote the robust conditional location functional solution of* (1.3) *and* (1.4) *when the vector* (X, Y) *has distribution* G.

Lemma 2.1 entails that the following result be a consequence of the almost everywhere weak convergence of $F_T(y|X = x)$ to $F(y|X = x)$ established in Theorem 4.1 of Section 4.

THEOREM 2.1. *Under* A1 *and* A2 *and* H1–H4 *we have that*:

(a) $g_T(x) \to g(x)$ *a.s. as* $T \to \infty$ *for almost all* $x(\mu)$.
(b) $g_T(x)$ *is asymptotically strongly robust* (ASR) *at* μ.

REMARK 2.4. Note that in Theorem 2.1 we do not impose any restriction on the probability distribution μ of the vector X. Hence the results obtained are robust and distribution-free in the sense that they are true for all μ.

3. Asymptotic results for robust nonparametric estimates based on nearest neighbor kernel weights. For the case of the k-nearest neighbor kernel methods we will require instead of H2, H3 and H4 the following assumptions.

H2′. The vector X has a density $f(x)$. $K\colon R^p \to R$ is a bounded nonnegative function, $\int K(u)\,du = 1$ and either of the following holds:

(a) $K(u) \le c_1 I_{\{\|u\| \le r\}}(u)$.
(b) f is bounded and $\int K^{2+\theta}(u)\,du < \infty$, where $\theta = 0$ if H1(a) holds and $\theta = 2\delta/(1-\delta)$ if H1(b) holds.

H3′. The sequence $\{k_T\colon t \in N\}$ satisfies $k_T \to \infty$ and $k_T/T \to 0$ as $T \to \infty$.

H4′. (a) For φ-mixing processes we require: The sequences $\{k_T\colon T \in N\}$ and $\{\varphi(n)\colon n \in N\}$ are such that $k_T/(n_T \log T) \to \infty$ as $T \to \infty$, where n_T is as in H4(a).
(b) In the α-mixing case we will require: There exists $\delta > 0$ such that $k_T^{(1+\delta)/4}/(T^{\delta/4} \log T) \to \infty$ as $T \to \infty$.

H5′. $K(uz) \ge K(z)$ for all $u \in (0,1)$.

REMARK 3.1. Remark 2.2(b) holds changing Th_T^p by k_T. Then for the uniform *k-NN* weights in the i.i.d. case we obtain the necessary and sufficient conditions on k_T obtained by Devroye (1982).

THEOREM 3.1. *Under* A1, A2, H1 *and* H2′–H5′ *we have that*:

(a) $\hat{g}_T(x) \to g(x)$ *a.s. as* $T \to \infty$ *for almost all* $x(\mu)$.
(b) $\hat{g}_T(x)$ *is ASR at* μ.

As in Section 2 this result follows from Lemma 2.1 and the almost everywhere weak convergence of $\hat{F}_T(y|X=x)$ to $F(y|X=x)$ established in Theorem 4.2.

COROLLARY 3.1. *Under* A1, A2, H1 *and* H2′–H5′ *the solution* $\hat{g}_T(x)$ *of*

$$\frac{1}{k_T} \sum_{t=1}^{k_T} \psi\left(\frac{Y_{R_t} - \hat{g}_T(x)}{\hat{s}_T(x)} \right) = 0,$$

i.e., the robust uniform k-NN estimator, is ASR and strongly consistent for almost all x.

REMARK 3.2. If $F(y|X=x)$ is a continuous function of x and f is continuous and bounded or f is continuous and K has a compact support, then it is easy to see that the results in Theorems 2.1 and 3.1 hold for all x.

4. Estimating the conditional distribution function. In this section we will study the strong consistency of $F_T(y|X=x)$ and $\hat{F}_T(y|X=x)$, defined by (2.4) and (2.5), respectively, to $F(y|X=x)$. Collomb (1984), Theorem 1 proves the complete convergence of $F_T(y|X=x)$ at each continuity point x of $F(y|X=x)$ for φ-mixing processes requiring some regularity conditions on the marginal density of X. In Theorem 5.1 we show the complete convergence of $F_T(y|X=x)$ for almost all x, for φ- and α-mixing processes under no assumptions whatsoever on the distribution of the vector (X, Y). In this sense it can be considered a distribution-free result. In Theorem 5.2 we show the complete convergence of $\hat{F}_T(y|X=x)$ under some weak regularity conditions stated in assumption H2′.

Given a Borel set $A \subset R$ we denote by $\phi_T(x)$, $\hat{\phi}_T(x)$ and $\phi(x)$ the functions

$$\phi_T(x) = \sum_{t=p+1}^{T} w_{tT}(x) I_A(Y_t),$$

$$\hat{\phi}_T(x) = \sum_{t=p+1}^{T} \hat{w}_{tT}(x) I_A(Y_t),$$

$$\phi(x) = E(I_A(Y)|X=x),$$

where w_{tT} and $\hat{w}_{tT}$ are defined in (1.5) and (1.6), respectively. In this section, h will stand for h_T.

We will use the following lemma due to Collomb [(1984), Lemma 1] which gives a sharper bound than a similar result obtained by Bosq (1975).

LEMMA 4.1 (Bernstein inequality for φ-mixing processes). *Let $\{\Delta_i\}$ be a sequence of φ-mixing random variables satisfying $E(\Delta_i) = 0$, $|\Delta_i| \le d$, $E(|\Delta_i|) \le \delta$ and $E(\Delta_i^2) \le D$. Denote $\tilde{\varphi}(m) = \sum_{i=1}^{m} \varphi(i)$ for each $m \in N$. Then for each $a > 0$ and $n \in N$ we have*

$$P\left(\left|\sum_{i=1}^{n} \Delta_i\right| > a\right) \le C_1 \exp\left(-\alpha a + \alpha^2 n C_2\right),$$

where $C_1 = 2\exp(3e^{1/2} n\varphi(m)/m)$, $C_2 = 6(D + 4\delta D\tilde{\varphi}(m))$ and where α and m are, respectively, a positive real number and a positive integer less than or equal to n satisfying $\alpha m d \le \frac{1}{4}$.

For the case of α-mixing processes we will use the following exponential inequality due to Doukhan, Leon and Portal (1984), Theorem 6.

LEMMA 4.2 (Bernstein inequality for α-mixing processes). *Let $\{\Delta_i\}$ be a sequence of geometrically α-mixing random variables satisfying $E(\Delta_i) = 0$ and $|\Delta_i| \le 1$. Given $0 < \delta < 1$, denote $\gamma = 2/(1-\delta)$ and $\sigma = \sup\{\|\Delta_i\|_\gamma, i \in N\}$ where $\|\Delta_i\|_\gamma^\gamma = E(|\Delta_i|^\gamma)$. Then there exist constants C_1 and C_2 which depend only on the mixing coefficients, such that*

$$P\left(\left|\sum_{i=1}^{n} \Delta_i\right| > a\right) \le \frac{C_1}{\delta} \exp\left(\frac{-C_{2,n} a^{1/2}}{n^{1/4}\sigma^{1/2}}\right),$$

where $C_{2,n} = C_2$ if $\sqrt{n}\,\sigma \le 1$ and $C_{2,n} = C_2 n^{1/4}\sigma^{1/2}$ if $\sqrt{n}\,\sigma > 1$.

4.1. *Consistency of the conditional distribution function for kernel weights.* The following result from real analysis will be needed. It is proved in Greblicki, Krzyżak and Pawlak [(1984), Lemma 1]. We will denote by μ the probability measure of the vector X.

LEMMA 4.3. *Let K satisfy* H2 *and g be an integrable function. Then*

$$\int K\left(\frac{x-y}{h}\right) g(y)\mu(dy) \Big/ \int K\left(\frac{x-y}{h}\right)\mu(dy) \to g(x) \quad \textit{as } h \to 0$$

for almost all $x(\mu)$.

When $K(t) = I_{\{\|u\| \le 1\}}(t)$ this result may be found, for instance, in Wheeden and Zygmund (1977), page 189.

Throughout this section S_r will be the closed ball of radius r centered at x.

THEOREM 4.1. *Assume* H1–H4. *Then*:

(i) $\phi_T(x) \to \phi(x)$ *a.s. for almost all x.*

(ii) $\lim_{T\to\infty} \sup_y |F_T(y|X=x) - F(y|X=x)| = 0$ *a.s. for almost all x.*

PROOF. (i) Let

$$a_T = E(K((X-x)/h)),$$
$$b_T = E(K((X-x)/h)\phi(X)),$$
$$\eta_{tT} = I_A(Y_t)K((X_t - x)/h)/a_T$$

and

$$\xi_{tT} = K((X_t - x)/h)/a_T.$$

Then

$$\phi_T(x) = \left(\frac{1}{T}\sum_{t=p+1}^{T} (\eta_{tT} - E(\eta_{tT})) + \left(1 - \frac{p}{T}\right)\frac{b_T}{a_T}\right)$$
$$\div \left(\frac{1}{T}\sum_{t=p+1}^{T} (\xi_{tT} - E(\xi_{tT})) + 1 - \frac{p}{T}\right).$$

By Lemma 4.3, $b_T/a_T \to \phi(x)$ as $T \to \infty$ for almost all $x(\mu)$, so it is enough to show that

$$\frac{1}{T}\sum_{t=p+1}^{T} (\eta_{tT} - E(\eta_{tT})) = \frac{S_T}{T} \to 0 \quad \text{a.s. as } T \to \infty \tag{4.1}$$

and

$$\frac{1}{T}\sum_{t=p+1}^{T} (\xi_{tT} - E(\xi_{tT})) \to 0 \quad \text{a.s. as } T \to \infty. \tag{4.2}$$

If (4.1) is proved, (4.2) follows by taking $A = \mathbb{R}$. We will consider separately the φ- and the α-mixing cases.

Denote by $\Delta_i = (\eta_{iT} - E(\eta_{iT}))/T$. Then we have $E(\Delta_i) = 0$ and $|\Delta_i| \le c_1(Ta_T)^{-1} = d$ where $c_1 = \sup_u K(u)$.

(a) Suppose that (X_t, Y_t) is a φ-mixing process satisfying H4(a). Then as $E(\Delta_i^2) \le c_1(T^2a_T)^{-1} = D$ and $E|\Delta_i| \le 2T^{-1} = \delta$, by Lemma 4.1 we have

$$P(|S_T/T| > \varepsilon) \le c_T \exp\big(-\alpha\varepsilon + \alpha^2 TC_T\big),$$

where $c_T = 2\exp(3\sqrt{e}\,T\varphi(n)/n)$, $C_T = 6(T^2a_T)^{-1}c_1(1 + 8\tilde{\varphi}(n))$ and $\alpha n c_1(Ta_T)^{-1} \le \frac{1}{4}$, $\alpha > 0$ and $n \in N$. Then, if we choose $\alpha = \beta Ta_T/(nc_1)$ and $\beta = \frac{1}{8}$ we get

$$P(|S_T/T| > \varepsilon) \le c_T \exp(-\beta Ta_T g(\varepsilon, T, n)/(nc_1)),$$

where $g(\varepsilon, T, n) = \varepsilon - \alpha TC_T$.

As in Lemma 2 of Collomb (1984) it is easy to see that $g(\varepsilon, T, n) > \varepsilon \min(\frac{1}{2}, c_1/2)$ and $c_T \le c$ for all $T \ge T_0$. Therefore

$$P(|S_T/T| > \varepsilon) \le c\exp(-\varepsilon Ta_T/16n) \le c\exp(-\varepsilon a T\mu(S_{rh})/16n)$$

by H2.

Then as $h^p r^p/\mu(S_{rh}) \to (d\lambda_1/d\mu)(x)$, where λ_1 is the μ-absolutely continuous part of the Legesgue measure on $\mathbb{R}^p$, H4(a) implies the complete convergence of S_T/T by taking $n = n_T$ if $n_T \to \infty$ or $n = n_1 > n_0$ if $n_T = n_0$ for $T \ge T_0$.

(b) Suppose now that (X_t, Y_t) is an α-mixing process verifying H4(b). Let $\gamma = 2/(1-\delta) > 2$. Then we have that

$$\begin{aligned} E(|\Delta_i|^\gamma) &= \frac{1}{T^\gamma} E(|\eta_{iT} - E(\eta_{iT})|^\gamma) \\ &\le \frac{2^{\gamma-1}}{T^\gamma} E\big(|\eta_{iT}|^\gamma + E(\eta_{iT})^\gamma\big) \le \frac{2^{\gamma-1}}{(Ta_T)^\gamma}\left[E\left(K^\gamma\left(\frac{X_T - x}{h}\right)\right) + a_T^\gamma\right] \\ &\le 2^\gamma \frac{E\big(K^\gamma((X_t - x)/h)\big)}{(Ta_T)^\gamma} \le \frac{2^\gamma c_1^{\gamma-1} a_T}{(Ta_T)^\gamma}. \end{aligned}$$

Then $T^{1/2}\sigma \le 2c_1^{1-1/\gamma}/(T^{1/2}a_T^{1-1/\gamma})$ where $\sigma = \sup\{E(|\Delta_i|^\gamma)^{1/\gamma}, i \in N\}$ and as by H4(b), $\lim_{T\to\infty} 2c_1^{1-1/\gamma}/(a_T^{1-1/\gamma}T^{1/2}) = 0$, by Lemma 5.2 we have

$$\begin{aligned} P\left(\left|\frac{S_T}{T}\right| > \varepsilon\right) &\le \frac{C_1}{\delta}\exp\big(-C_2\varepsilon^{1/2}\big(T^{1/4}a_T^{(1+\delta)/4}\big)\big) \\ &\le \frac{C_1}{\delta}\exp\big(-C_3 T^{1/4}\mu(S_{rh})^{(1+\delta)/4}\big) \end{aligned}$$

and the desired result follows from H4(b) as $h^p r^p/\mu(S_{rh}) \to (d\lambda_1/d\mu)(x)$.

(ii) Follows from (i) by an argument similar to the one used to prove the Glivenko–Cantelli theorem. □

REMARK 4.1. If $A = \mathbb{R}$ and $K(t) = I_{\{\|t\| \le r\}}(t)$ we have $\|\eta_{iT}/T\|_\gamma \le \sigma + E(\eta_{iT})/T = \sigma + (1/T)$. Then as $\|\eta_{iT}/T\|_\gamma = (Ta_T^{1-1/\gamma})^{-1}$ we have that

$[T^{1/4}\alpha_T^{(1+\delta)/4}]^{-1} \le [T^{1/2}\sigma + T^{-1/2}]^{1/2}$. Therefore, the required convergence rate of the window bandwidth h for α-mixing processes is sharp in the sense that it is the best possible rate that can be obtained from the exponential inequality given in Lemma 5.2.

4.2. *Consistency of the conditional distribution function for nearest neighbor kernel weights.* In this section we will also require that the marginal distribution of the vector X has a density $f(x)$. Lemma 4.5 also shows that Theorem 4.1 holds without requiring H2.

LEMMA 4.4. *Under* H2′ *and* H3 *we have that for any Borel set* $A \subset R$,

$$h^{-p}E\big(K\big((X_{p+1} - x)/h\big)I_A(Y_{p+1})\big) \to f(x)\phi(x) \quad \text{as } h \to 0$$

for almost all x.

PROOF. Using Lemma 4.3 and H2′, it is easy to see that the problem can be reduced to considering the case where $A = R$, i.e., it is enough to show that

$$\lim_{h\to 0} h^{-p}E[K((X - x)/h)] = f(x) \quad \text{for almost all } x. \tag{4.3}$$

When f is bounded and continuous at x, (4.3) follows easily from the dominated convergence theorem. Let now f be any bounded measurable function $\int f(x)\,dx = 1$ and g_k be a sequence of nonnegative continuous and bounded functions such that $\lim_{k\to\infty} \int |f(z) - g_k(z)|\,dz = 0$ and $\sup_x g_k(x) \le M$ for all $k \ge k_0$.

For each fixed k we have

$$\begin{aligned} L &= \limsup_{h\to 0} \left| h^{-p} \int K((z - x)/h) f(z)\,dz - f(x) \right| \\ &\le \limsup_{h\to 0} \left| h^{-p} \int K((z - x)/h) g_k(z)\,dz - g_k(x) \right| + |g_k(x) - f(x)| \qquad (4.4) \\ &\quad + \limsup_{h\to 0} h^{-p} \int K((z - x)/h) |g_k(z) - f(z)|\,dz. \end{aligned}$$

As g_k is continuous and bounded, the first term on the right in (4.4) is zero. Consider first the case when $K(u) \le c_1 I_{\{\|u\| \le r\}}$.

Given $\varepsilon > 0$, let $E_\varepsilon = \{x\colon L > \varepsilon\}$. Then, an argument analogous to that of the Lebesgue differentiation theorem [see, for instance, Wheeden and Zygmund (1977), page 106] using the Hardy–Littlewood maximal function of $(f - g_k)$ and Chebyshev's inequality leads to

$$|E_\varepsilon|_e \le 2c_1^{-1}\varepsilon a(p) \int |f(y) - g_k(y)|\,dy + 2\varepsilon^{-1} \int |f(y) - g_k(y)|\,dy,$$

where $|\cdot|_e$ stands for the outer Lebesgue measure and $a(p)$ is a constant depending only on p. By letting $k \to \infty$ it follows that $|E_\varepsilon|_e = 0$ and therefore

$$h^{-p} \int K((z - x)/h) f(z)\, dz \to f(x) \quad \text{as } h \to 0 \text{ for almost all } x.$$

Finally suppose that H2′(b) holds. Given $\varepsilon > 0$ choose r such that $\int_{|u|>r} K(u)\, du < \varepsilon/4M$. Then

$$h^{-p} \int K((z - x)/h) |g_k(z) - f(z)|\, dz \le h^{-p} c_1 \int_{S_{rh}} |g_k(z) - f(z)|\, dz$$

$$+2M \int_{|u|>r} K(u)\, du < c_1 (g_k - f)^*(x) + \varepsilon/2$$

and the proof follows as above. □

LEMMA 4.5. *Assume that H1, H3, H4 and H2′ hold. Then for any Borel set $A \subset R$ we have that*

$$\frac{1}{Th^p} \sum_{t=p+1}^{T} K((X_t - x)/h) I_A(Y_t) \to f(x)\phi(x) \quad a.s. \text{ as } T \to \infty$$

for almost all x.

PROOF. Denote by $\eta_{tT} = h^{-p} K((X_t - x)/h) I_A(Y_t)$. Then by Lemma 4.4 it suffices to show that

$$\frac{1}{T} \sum_{t=p+1}^{T} (\eta_{tT} - E(\eta_{tT})) = \frac{S_T}{T} \to 0 \quad \text{a.s. as } T \to \infty.$$

We will use again Lemmas 4.1 and 4.2 for φ- and α-mixing processes, respectively, with $\Delta_i = (\eta_{iT} - E(\eta_{iT}))/T$.

By Lemma 4.4 we have

$$h^{-p} \int K^2((z - x)/h) f(z)\, dz \to f(x) \int K^2(u)\, du \quad \text{as } T \to \infty$$

and

$$h^{-p} \int K((z - x)/h) f(z)\, dz \to f(x) \quad \text{as } T \to \infty$$

for almost all x.

Then

$$E(\Delta_i^2) \le (T^2 h^p)^{-1} h^{-p} E\big(K^2((X_{p+1} - x)/h)\big)$$

$$\le (T^2 h^p)^{-1} 2 f(x) \int K^2(u)\, du = D,$$

$$E(|\Delta_i|) \le 2T^{-1} h^{-p} E\big(K((X_{p+1} - x)/h)\big) \le 4 f(x) T^{-1} = \delta$$

for $T \ge T_0$ and $|\Delta_i| \le (Th^p)^{-1} \sup_u K(u) = d$. For φ-mixing processes the proof follows as in Theorem 4.1 from Lemma 4.1.

Finally, using Lemma 4.4 we obtain

$$\sigma = E(|\Delta_i|^\gamma)^{1/\gamma} \le \frac{1}{T}\left[(E(\eta_{iT}^\gamma))^{1/\gamma} + E(\eta_{iT})\right]$$
$$\le \frac{1}{T}\left[(h^p)^{(1-\gamma)/\gamma}(2f(x))^{1/\gamma}\left(\int K^\gamma(u)\,du\right)^{1/\gamma} + 2f(x)\right] \le T^{-1}(h^p)^{(1-\gamma)/\gamma}C^*$$

and Lemma 4.2 implies the desired result as in Theorem 4.1. □

Lemma 4.6 proved in Boente and Fraiman (1988) shows that the complete consistency of the density estimators proposed by Loftsgaarden and Quesenberry (1965) and studied by Wagner (1973) still holds for φ- and α-mixing processes.

This result will be used in proving the complete consistency of $\hat{F}_T(y|X = x)$.

We will denote by λ the Lebesgue measure on $\mathbb{R}^p$ and by V_r the closed ball of radius r centered at 0.

LEMMA 4.6. *Let $\{X_t: t \ge p + 1\}$ be a strictly stationary process with density $f(x)$ satisfying* H1. *Define $H_T^p = \|X_{R_k} - x\|^p$ where $k = k_T$ verifies* H3′ *and* H4′. *Then*

$$f_T(x) = k_T\big(T\lambda(V_{H_T})\big) \to f(x) \quad \textit{completely, for almost all } x \textit{ as } T \to \infty.$$

The proof may be found in Boente and Fraiman (1988).

The following lemma can be found in Collomb (1980), page 162.

Let (X_t, B_t), $1 \le t \le T$, be a sequence of random vectors $X_t \in R^p$, $B_t \in R^+$ and k: $R \times R^p \to R^+$ a measurable function such that

$$u \le u' \Rightarrow k(u, z) \le k(u', z)\,\forall z \in \mathbb{R}^p. \tag{4.5}$$

Denote by $c_T(D) = \sum_{t=1}^T B_t k(D, X_t)/\sum_{t=1}^T k(D, X_t)$.

LEMMA 4.7. *Let $(D_T)_{T\in N}$ be a sequence of random variables. If for all β, $0 < \beta < 1$, there exist two sequences $(D_T^-)_{T\in N}$ and $(D_T^+)_{T\in N}$ satisfying*

$$D_T^- \le D_T^+ \quad \forall T \quad \textit{and} \quad I_{\{D_T^- \le D_T \le D_T^+\}} \to 1 \quad a.s., \tag{4.6}$$

$$c_T(D_T^-) \to c \quad \textit{and} \quad c_T(D_T^+) \to c \quad a.s., \tag{4.7}$$

$$\sum_{t=1}^T k(D_T^-, X_t)\bigg/\sum_{t=1}^T k(D_T^+, X_t) \to \beta \quad a.s., \tag{4.8}$$

then $c_T(D_T) \to c$.

THEOREM 4.2. *Under* H1 *and* H2′–H5′ *we have that*:

(i) $\hat{\phi}_T(x) \to \phi(x)$ *a.s. as $T \to \infty$ for almost all x.*
(ii) $\lim_{T\to\infty} \sup_{y\in\mathbb{R}}|\hat{F}_T(y|X = x) - F(y|X = x)| = 0$ *a.s. for almost all x.*

PROOF. (i) Take in Lemma 4.7, $B_t = I_A(Y_t)$, $D_T = H_T$ and $k(u, z) = K((z - x)/u)$. For all x such that $f(x) > 0$ denote $h_T = [k_T/(Tf(x)\lambda(V_1))]^{1/p}$.

Then h_T satisfies H3 and H4. Given $\beta \in (0,1)$ define $D_T^- = D_T^-(\beta) = h_T\beta^{1/2p}$ and $D_T^+ = D_T^+(\beta) = h_T\beta^{-1/2p}$.

The proof follows as in Collomb (1980), Proposition 2 using Lemmas 4.5 and 4.6.

(ii) Follows as in Theorem 4.1. □

REFERENCES

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

BOENTE, G. and FRAIMAN, R. (1988). Consistency of a nonparametric estimate of a density function for dependent variables. *J. Multivariate Anal.* **25** 90–99.

BOENTE, G. and FRAIMAN, R. (1989). Robust nonparametric regression estimation. *J. Multivariate Anal.* **29** 180–198.

BOENTE, G., FRAIMAN, R. and YOHAI, V. (1987). Qualitative robustness for stochastic processes. *Ann. Statist.* **15** 1293–1312.

BOSQ, D. (1975). Inégalité de Bernstein pour les procéssus stationnaires et mélangeant. Applications. *C. R. Acad. Sci. Paris Ser. A* **281** 1095–1098.

BOSQ, D. (1980). Une méthode non parametrique de prédiction d'un processus stationnaire. Prédictions d'une mesure aléatoire. *C. R. Acad. Sci. Paris Ser. A* **290** 711–713.

BOX, G. and JENKINS, G. (1970). *Time Series Analysis, Forecasting and Control.* Holden-Day, San Francisco.

CARBON, M. (1983). Bernstein's inequality for strong mixing non-stationary processes. Applications. *C. R. Acad. Sci. Paris Sér. I Math.* **297** 303–306.

COLLOMB, G. (1980). Estimation de la régression par la méthode des k points les plus proches avec noyau: Quelques propriétés de convergence ponctuelle. *Statistique non paramétrique asymptotique. Lecture Notes in Math.* **821** 159–175.

COLLOMB, G. (1981). Estimation non-paramétrique de la régression: Revue bibliographique. *Internat. Statist. Rev.* **49** 75–93.

COLLOMB, G. (1982). Prédiction non paramétrique: Étude de l'erreur quadratique du prédictogramme. *C. R. Acad. Sci. Paris Sér. I Math.* **294** 59–62.

COLLOMB, G. (1984). Propiétés de convergence presque complete du predicteur a noyau. *Z. Wahrsch. verw. Gebiete* **66** 441–460.

COLLOMB, G. and HÄRDLE, W. (1984). Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations. *Stochastic Process. Appl.* **23** 77–89.

COX, D. (1981). Metrics on stochastic processes and qualitative robustness. Technical Report 3, Dept. Statistics, Univ. Washington.

DEVROYE, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.* **9** 1310–1319.

DEVROYE, L. (1982). Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Z. Wahrsch. verw. Gebiete* **61** 467–481.

DOUKHAN, P. and GHINDES, M. (1980). Estimations dans le processus "$X_{n+1} = f(X_n) + \varepsilon_n$." *C. R. Acad. Sci. Paris Ser. A* **291** 61–64.

DOUKHAN, P. and GHINDES, M. (1983). Estimation dans le processus autorégressif d'ordre 1. *Stochastic Process. Appl.* **15** 271–293.

DOUKHAN, P., LEON, J. and PORTAL, F. (1984). Vitesse de convergence dans le théoreme central limite pour des variables aléatoires mélangeantes a valeurs dans un espace de Hilbert. *C. R. Acad. Sci. Paris Sér. I Math.* **298** 305–308.

DOUKHAN, P., LEON, J. and PORTAL, F. (1985). Principes d'invariance faible pour la measure empirique d'une suite de variables aléatoires mélangeante. Unpublished.

GREBLICKI, W., KRZYŻAK, A. and PAWLAK, M. (1984). Distribution-free pointwise consistency of kernel regression estimate. *Ann. Statist.* **12** 1570–1575.

HÄRDLE, W. (1984). Robust regression function estimation. *J. Multivariate Anal.* **14** 169–180.

HÄRDLE, W. and TSYBAKOV, A. B. (1988). Robust nonparametric regression with simultaneous scale curve estimation. *Ann. Statist.* **16** 120–135.

IBRAGIMOV, I. A. and LINNIK, YU. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

LOFTSGAARDEN, D. O. and QUESENBERRY, C. D. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.* **36** 1049–1051.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

NGUYEN, H. T. and PHAM, D. T. (1981). Nonparametric estimation in diffusion model by discrete sampling. *Publ. Inst. Statist. Univ. Paris* **26**(2) 89–109.

PAPANTONI-KAZAKOS, P. and GRAY, R. M. (1979). Robustness of estimators on stationary observations. *Ann. Probab.* **7** 989–1002.

PELIGRAD, M. (1986). Properties of uniform consistency of the kernel estimators of density and of regression functions under dependence assumptions. Unpublished.

PHAM, D. T. (1981). Nonparametric estimation of the drift coefficient in diffusion equation. *Math. Operationsforsch. Statist. Ser. Statist.* **12** 61–74.

ROBINSON, P. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal.* **4** 185–207.

ROBINSON, P. (1984). Robust nonparametric autoregression. *Robust and Nonlinear Time Series Analysis. Lectures Notes in Statist.* **26** 247–255.

ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U.S.A.* **42** 43–47.

ROUSSAS, G. (1969). Nonparametric estimation of the transition distribution function of a Markov process. *Ann. Math. Statist.* **40** 1386–1400.

STONE, C. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

TSYBAKOV, A. B. (1982). Robust estimates of a function. *Problems Inform. Transmission* **18** 190–201.

WAGNER, T. J. (1973). Strong consistency of a nonparametric estimate of a density function. *IEEE Trans. Systems Man Cybernet.* **3** 289–372.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.

WHEEDEN, R. L. and ZYGMUND, A. (1977). *Measure and Integral*. Dekker, New York.

YAKOWITZ, S. J. (1985). Nonparametric density estimation, prediction and regression for Markov sequences. *J. Amer. Statist. Assoc.* **80** 215–221.

DEPARTAMENTO DE MATEMATICAS
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
CIUDAD UNIVERSITARIA, PAB. NO. 1
BUENOS AIRES 1428
ARGENTINA